

UNIVERSITY OF MASSACHUSETTS
 Department of Mathematics and Statistics
 Advanced Exam Version I
 Thursday, August 30, 2018

Work all problems. 60 points are needed to pass at the Masters Level and 75 to pass at the Ph.D. level.

1. Consider the linear model

$$\mathbf{y} = \begin{bmatrix} \mathbf{1}_5 & 0 & 0 \\ 0 & \mathbf{1}_5 & 0 \\ 0 & 0 & \mathbf{1}_5 \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} + \epsilon$$

where $\mathbf{1}_5 = (1, 1, 1, 1, 1)'$ and

$$\mathbf{y}' = [y_{11}, y_{12}, y_{13}, y_{14}, y_{15}, y_{21}, y_{22}, y_{23}, y_{24}, y_{25}, y_{31}, y_{32}, y_{33}, y_{34}, y_{35}].$$

- (a) Show that the least square estimate of b_i is $\hat{b}_i = \bar{y}_{i.} = \sum_{j=1}^5 y_{ij}/5, i = 1, 2, 3$.
- (b) Show that residual sum of squares due to regression is $SSR_m = \frac{y_{1.}^2}{5} + \frac{y_{2.}^2}{5} + \frac{y_{3.}^2}{5} - \frac{y_{..}^2}{15}$, where $y_{i.} = \sum_{j=1}^5 y_{ij}$ and $y_{..} = \sum_{i=1}^3 y_{i.}$
- (c) For the hypothesis $H : b_1 - b_2 = 0$, find the F -statistic and the estimate of \mathbf{b} under H .
2. Suppose an oil company gets its crude oil from four different sources, refines it in three different refineries using the same two processes in each refinery. In one part of the refining process, a measurement of efficiency is taken as a percentage and recorded as an integer between 0 and 100. The following table shows the available measurement of efficiency for different samples of oil.

Refinery	Process	Source			
		Texas	Oklahoma	Gulf of Mexico	Iran
Galveston	1	31, 33, 44, 36	38	26	—
	2	37, 59	42	—	—
Newark	1	—	—	42	34, 42, 28
	2	39	36	32, 38	—
Savannah	1	42	36	—	22
	2	—	42, 46	26	37, 43

Write down the linear model $\mathbf{y} = \mathbf{X}\mathbf{b} + \epsilon$ giving the explicit forms of \mathbf{X} , \mathbf{b} , and ϵ for each of the following scenarios:

- (a) Using the eight observations on Texas, consider only the effects of refinery and process on efficiency.
- (b) Using the eight observations on Texas, include interactions between refinery and process.
- (c) Using the eight observations on Texas, consider the effects of refinery and process on efficiency and assume that process is nested within refinery.
- (d) For all 25 observations, write down the equation of the linear model for considering the effect of source, refinery, and process on efficiency. Do not include interactions.
3. Observed data are $\{y_i, x_{i1}, x_{i2}\}_{i=1}^n$, and $x_{i1} + x_{i2} = 1$ for all i , and $n > 3$. Let $\mathbf{X}_{\mathbf{k}} = \begin{pmatrix} 1 & x_{1k} \\ \vdots & \vdots \\ 1 & x_{nk} \end{pmatrix}$ and suppose $\mathbf{X}_{\mathbf{k}}$ has rank 2 for $k = 1, 2$. (In all parts you can cite known results without proof as part of your answers.)

- (a) Model 1 is $y_i = \beta_0 + \beta_1 x_{i1} + \varepsilon_i$, and Model 2 is $y_i = \gamma_0 + \gamma_1 x_{i2} + \varepsilon_i$. Prove or disprove: \hat{y}_i s are the same for models 1 and 2.
- (b) Is there a relationship between $\hat{\beta}_1$ and $\hat{\gamma}_1$? If so, what is it and why?
- (c) If models 1 and 2 do not include intercept terms, does that change your answer to part (a)? Why or why not?
- (d) Next, Model 3 is $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i1}^2 + \varepsilon_i$, and Model 4 is $y_i = \gamma_0 + \gamma_1 x_{i2} + \gamma_2 x_{i2}^2 + \varepsilon_i$. Prove or disprove: \hat{y}_i s are the same for models 3 and 4.
4. Suppose observed data are $\{y_{i1}, y_{i2}\}_{i=1}^n$, and consider the model $y_{ik} = \beta_i + \varepsilon_{ik}, \varepsilon_{ik} \sim iid N(0, \sigma^2), k = 1, 2$.
- (a) Find the MLEs of β_i s and σ^2 .
- (b) Are the MLEs consistent or not? Why or why not? (Note, this is as $n \rightarrow \infty$.)
5. Let X be the age in days at which a car stops working. One model for the distribution of X is $X \sim exponential(\lambda)$. The pdf is $f(x, \lambda) = \lambda \exp(-x\lambda), x \geq 0, \lambda > 0$.
- (a) Suppose the car is working after 250 days. Derive the expected value for the number of additional days that the car will work.
- (b) Suppose the car is working after x days. Show that the probability that it fails in the interval $(x, X + \epsilon)$ is approximately $\epsilon\lambda$ for small $\epsilon > 0$.
- (c) What do your answers for part (a) and (b) say about whether the exponential model is sensible or not?